New Monte Carlo Algorithm for Protein Folding

Helge Frauenkron,[1] Ugo Bastolla,[1] Erwin Gerstner,[1,2] Peter Grassberger,[1,2] and Walter Nadler[1]
[1]*HLRZ c/o Forschungszentrum Jülich, D-52425 Jülich, Germany*
[2]*Physics Department, University of Wuppertal, D-42097 Wuppertal, Germany*
(Received 14 May 1997)

We demonstrate that the recently introduced pruned-enriched Rosenbluth method leads to extremely efficient algorithms for the folding of simple model proteins. We test them on several models for lattice heteropolymers, and compare the results to published Monte Carlo studies. In all cases our algorithms are faster than previous ones, and in several cases we find new minimal energy states. In addition, our algorithms give estimates for the partition sum at finite temperatures. [S0031-9007(98)05636-1]

PACS numbers: 87.15.By, 02.70.Lq, 87.10.+e

Protein folding [1] is one of the outstanding problems in mathematical biology. It is concerned with the problem of how a given sequence of amino acids assumes precisely that geometrical shape which is biologically useful. Currently, it is much easier to find coding DNA (and, thus, amino acid) sequences than to find the corresponding structures. Thus, solving the folding problem would be a major breakthrough in understanding the biochemistry of the cell, and in designing artificial proteins.

In this Letter we are concerned only with the most straightforward direct approach: Given a sequence, a molecular potential, and no other information, find the ground state and the equilibrium state at physiological temperatures. Note that we are not concerned with the kinetics of folding, but only in the final outcome. Also, we will not address the problems of how to find good molecular potentials, and what is the proper level of detail in describing proteins. Instead, we will use simple coarse-grained models which have been proposed in the literature and have become standards in testing the efficiency of folding algorithms.

The models we study are heteropolymers which live on $3d$ or $2d$ regular lattices. They are self-avoiding chains with attractive or repulsive interactions between neighboring nonbonded monomers. These interactions can have continuous distributions [2], but the majority of authors considered only two types of monomers. In the HP model [3,4] they are hydrophobic (H) and polar (P), with $(\epsilon_{HH}, \epsilon_{HP}, \epsilon_{PP}) = -(1, 0, 0)$. Since this leads to highly degenerate ground states, alternative models were proposed, e.g., $\vec{\epsilon} = -(3, 1, 3)$ [5] and $\vec{\epsilon} = -(1, 0, 1)$ [6].

The algorithms we apply here are variants of the pruned-enriched Rosenbluth method (PERM) [7], a chain growth algorithm based on the Rosenbluth-Rosenbluth (RR) [8] method. Monomers are added sequentially, the nth monomer being placed at site i with probability $p_n(i)$. In *simple sampling*, $p_n(i)$ is uniform on all neighbors of the last monomer, leading to exponential attrition. The original RR method avoids this by using a uniform $p_n(i)$ on all *vacant* neighbors of i_{n-1}. More generally, we call any nonuniform choice of $p_n(i)$ a generalized RR method. The relative thermal weight of a particular chain conformation of length n is then determined by $W_n = m_n \exp(-\beta \Delta E_n) W_{n-1}$, with $W_1 = 1$; ΔE_n is the energy gain from adding monomer n; and m_n is the Rosenbluth factor, $m_n = \sum_{j \in \{nn\}} p_n(j)/p_n(i)$. We note that W_n is also an estimate for the partition function Z_n of the n-monomer chain [7]. Chain growth is stopped when the final size is reached and started anew from $n = 1$.

In easy cases, $p_n(i)$ can be chosen so that Boltzmann and Rosenbluth factors—or Rosenbluth factors for different n—cancel, leading to narrow weight distributions. But, in general, this algorithm produces a wide spread in weights that can lead to serious problems [9]. On the other hand, since the weights accumulate as the chain grows, one can interfere during the growth process by "pruning" conformations with low weights and enriching high-weight conformations. This is, in principle, similar to population based methods in polymer simulations [10,11] and in quantum Monte Carlo (MC) algorithms [12]. However, our implementation is different. Pruning is done stochastically: If the weight of a conformation has decreased below a threshold $W_n^<$, it is eliminated with probability $1/2$, while it is kept and its weight is doubled in the other half of cases. Enrichment [13] is done independently of this: If W_n increases above another threshold $W_n^>$, the conformation is replaced by n_c copies, each with weight W_n/n_c. Technically, this is done by putting onto a stack all information about conformations which still has to be copied. This is most easily implemented by recursive function calls [7]. Thereby the need for keeping large populations of conformations [10–12] is avoided. PERM has proven extremely efficient for studies of lattice homopolymers near the θ point [7], their phase equilibria [14], and of the ordering transition in semistiff polymers [15].

The main freedom when applying PERM consists in the *a priori* choice of the sites as to where to place the next monomer, i.e., the probabilities $p_n(j)$, in the thresholds $W_n^<$ and $W_n^>$ for pruning and enrichment, and in the number of copies n_c made upon enrichment. All of these features do not affect the correctness of the algorithm, but they can greatly influence its efficiency.

 © 1998 The American Physical Society

They may depend arbitrarily on chain lengths and on local conformations, and they can be changed freely at any time during the simulation. Thus, the algorithm can "learn" during the simulation [7].

In order to apply PERM to heteropolymers at very low temperatures, the strategies proposed in [7] are modified as follows.

(1) For homopolymers near the theta point it had been found that the best choice for the placement of monomers was not according to their Boltzmann weights, but uniformly on all allowed sites [7,14] such as in the original RR. This is due to cancellation between Boltzmann and RR factors: Larger Boltzmann factors correspond to higher densities and, thus, to smaller RR factors [9].

For a heteropolymer this has to be modified, as there is no longer a unique relationship between density and the Boltzmann factor. In a strategy of "anticipated importance sampling" we should preferentially place monomers in sites with mostly attractive neighbors. Assume that we have two types of monomers and we want to place a type-A monomer. If an allowed site has m_B neighbors of type B ($B = H, P$), we select this site with probability $\propto 1 + a_{AH} m_H + a_{AP} m_P$. Here, a_{AB} are constants with $a_{AB} > 0$ for $\epsilon_{AB} < 0$ and vice versa.

(2) Most naturally, the $W_n^>$ and $W_n^<$ are chosen proportional to the estimated partition sum Z_n (i.e., the average of the W_n already generated), e.g., $W_n^< = cZ_n$, $c \approx 0.5$, and $W_n^> = rW_n^<$, $r \approx 10$ [7]. But this becomes inefficient at very low T since Z_n will be underestimated as long as no low-energy state is found. But when this finally happens, $W_n^>$ is too small and, thus, too many (correlated) copies are produced. This costs CPU time but does not increase the quality of sampling.

This problem could be avoided by increasing $W_n^>$ and $W_n^<$ during particularly successful "tours" (a tour is the set of conformations derived from a single start [7]). But then also the average number of long chains is decreased in comparison with short chains. To reduce this effect and to create a bias towards a sample which is flat in chain length, we multiply by some power of M_n/M_1, where M_n is the number of generated chains of length n. With $\mathcal{N}(n)$ denoting the number of chains generated during the current tour we used $W_n^< = CZ_n[(1 + \mathcal{N}(n)/M)(M_n + M)/(M_1 + M)]^2$, with C a constant of order unity and M a constant of order 10^4–10^5.

(3) For the number of copies n_c created when W_n surpasses $W_n^>$, a good choice is int$[1 + \sqrt{W_n/W_n^>}]$.

(4) In some cases we did not start to grow the chain from one end but from a point in the middle. We grew first in one direction, and then in the other. Results were averaged over all possible starting points. The idea behind this is that real proteins have folding nuclei, and it should be most efficient to start from such a nucleus. In some cases this approach was very successful and speeded up the ground state search substantially, in others not.

(5) Special tricks were employed for "compact" configurations filling a square or a cube [16].

Let us now discuss our results. Items (a)–(c) concern the original HP model [3] with $\vec{\epsilon} = -(1, 0, 0)$.

(a) Ten sequences of length $N = 48$ were given in [17]. Each was designed by minimizing the energy of a particular target conformation under the constraint of constant composition. The authors tried to find the lowest-energy states with a heuristic MC method [18], and an exact enumeration of low-energy states [19] (which cannot be generalized to other models). The MC method failed in all but one case. Precise CPU times were not quoted. With PERM we succeeded in reaching ground states in *all* cases, average CPU time per sequence ranging from a few seconds to several hours (all times refer to a SUN ULTRA SPARC, 167 MHz). We verified also that these ground states are highly degenerate, and that there are no gaps between ground and first excited states. Thus, none of these sequences are good folders, though they were designed specifically for this purpose.

(b) In Ref. [20] two versions of a genetic algorithm were used to simulate $2d$ HP chains of lengths 20 to 64, and compared to other MC algorithms. Ground state energies were supposed to be known since the chains had been specially designed. In all cases we reached the ground state energies proposed by the authors in less than 1 h CPU time, except for the sequence of length 64. For that sequence we obtained $E = -39$, while none of the algorithms used in [20] reached energies below -37. For the chain with length 60, we found several states with $E = -36$ although the authors had claimed $E \geq -34$ by construction (see Table I).

(c) Two $2d$ HP chains with $N = 100$ were studied in [21]. The authors claimed that the native conformations are compact, fit exactly into a 10×10 square, and have energies -44 and -46. These energies were found by a specially designed MC algorithm which should be particularly efficient for compact conformations. We found noncompact (degenerate) conformations with energies -47 and -49 (see Table I), while our lowest-energy compact states (also degenerate) have $E = -46$ and -47 [16].

(d) Sequences with $N = 27$ and with continuous interactions were studied in Ref. [2]. Interaction strengths were sampled from Gaussians and were permuted to obtain good folders. In all cases we could reach the supposed ground state energies, within less than 1 h in the worst case. This time the design had been successful, and all sequences showed gaps between the ground state and the bulk of low-lying states. These gaps were in some cases filled by conformations which were very similar to the ground state, so that they could not prevent these sequences from being counted as good folders. In no case did we find energies lower than those quoted in [2].

(e) Short sequences with which we had neither problems nor surprises were given in several papers: 48-mers in [6], 27-mers in [22], and sequences with $N \leq 36$ in [19].

TABLE I. Newly found lowest energy states for binary sequences with interactions $\vec{\epsilon} = (\epsilon_{HH}, \epsilon_{HP}, \epsilon_{PP})$. Configurations are encoded as sequences of *r* (right),*l* (left), *u* (up), *d* (down), *f* (forward), and *b* (backward).

N	d	$\vec{\epsilon}$	Sequence *example conformation*	old $E_{\min}$ *our* $E_{\min}$	Ref.
60	2	$-(1,0,0)$	$P_2H_3PH_8P_3H_{10}PHP_3H_{12}P_4H_6PH_2PHP$ $r_5d_2lul_3dld_2(ru)_2rd_2ldldrdr_2uluru_2rd_2rdldr_2u_3lu_3rd_2rur$	-34 -36	[20]
100	2	$-(1,0,0)$	$P_6HPH_2P_5H_3PH_5PH_2P_2(P_2H_2)_2PH_5PH_{10}PH_2PH_7P_{11}H_7P_2HPH_3P_6HPH_2$ $r_6ur_2u_3rd_5luldl_2drd_2ru_2r_3(rulu)_2urdrd_2ru_3lur_3dld_2rur_5d_3l_5uldl_2d_3ru_2r_3d_3l_2urul$	-44 -47	[21]
100	2	$-(1,0,0)$	$P_3H_2P_2H_4P_2H_3(PH_2)_3H_2P_8H_6P_2H_6P_9HPH_2PH_{11}P_2H_3PH_2PHP_2HPH_3P_6H_3$ $ul_2drdl_2u_3ld_4ldrdl_2u_2l_2d_3l_2uru_3r_2u_3rd_3ru_4rul_5dldr_2d_2luldldrdldlu_3lul_2ulur_2dr_2u_3rd_4l$	-46 -49	[21]
80	3	$-(1,0,1)$	$PH_2P_3(H_3P_2H_3P_3H_2P_3)_3H_4P_4(H_3P_2H_3P_3H_2P_3)H_2$ $lbruflbl_2br_2drur_2dldl_3ulfrdr_3urfldl_3ulurur_3drblul_3br_3bl_3dldrdr_3urul_2dlu$	-94 -98	[6]

(f) The most interesting case is a two-species 80-mer with interactions $-(1,0,1)$ studied first in [6]. These particular interactions were chosen because it was hoped that they would lead to compact conformations. Indeed, the sequence was specially designed to form a "four helix bundle" which fits perfectly into a $4 \times 4 \times 5$ box (see Fig. 1). Its energy in this putative native state is -94. Although the authors of [6] used highly optimized codes, they were not able to recover this state by MC. Instead, they reached only $E = -91$. Supposedly, a different state with $E = -94$ was found in [21], but Fig. 10 of this paper, which is claimed to show this conformation, has a much higher value of E.

Even without much tuning, our algorithm gave $E = -94$ after a few hours, but it did not stop there. After a number of conformations with successively lower energies, the final candidate for the native state has $E = -98$. It again has a highly symmetric shape, although it does not fit into a $4 \times 4 \times 5$ box (see Fig. 2). It has twofold degeneracy (the central $2 \times 2 \times 2$ box in the front of Fig. 2 can be flipped), and both conformations were actually found in the simulations. Optimal parameters for the ground state search in this model are $\beta = 1/kT \approx 2.0$, $a_{PP} = a_{HH} \approx 2$, and $a_{HP} \approx -0.13$. With these, the average CPU times for finding $E = -94$ and $E = -98$ are about 20 min and 80 h, respectively [23].

A surprising result is that the monomers are arranged in four homogeneous layers in Fig. 2, while they had formed only three layers in the putative ground state of Fig. 1. Since the interaction should favor the segregation of different type monomers, one might have guessed that a conformation with a smaller number of layers should be favored. We see that this is outweighed by the fact that both monomer types can form large double layers in the new conformation. Again, our new ground state is not compact in the sense of minimizing the surface, and hence it disagrees with the widespread prejudice that native states are compact.

We also constructed histograms of the energy distribution. Combining them with similar histograms obtained at higher temperatures [24], we obtained average

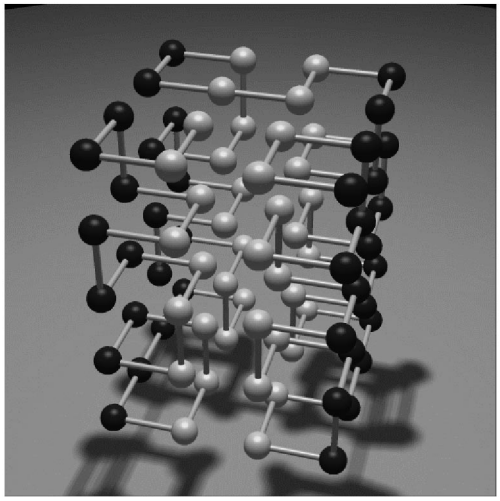

FIG. 1. Putative native state of the four helix bundle sequence, as proposed in [6]. It has $E = -94$, fits into a rectangular box, and consists of three homogeneous layers. Structurally, it can be interpreted as four helix bundles.

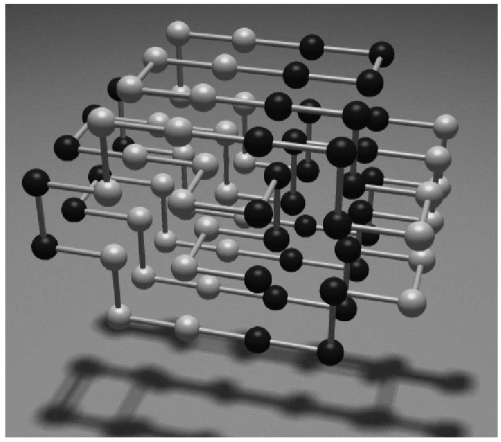

FIG. 2. Conformation of the four helix bundle sequence with $E = -98$. We propose that this is the actual ground state. Its shape is highly symmetric although it does not fit into a rectangular box. It is not degenerate, except for a flipping of the central front $2 \times 2 \times 2$ box.

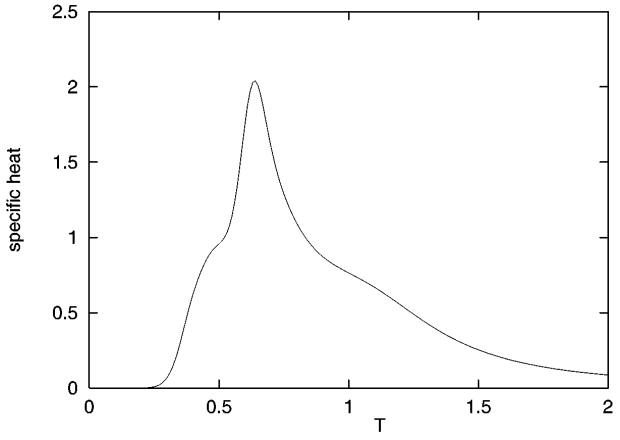

FIG. 3. Specific heat (heat capacity per monomer) of the 80-mer four helix bundle vs T.

energies and heat capacities. The specific heat (Fig. 3) shows a large peak at $T = 0.62$ and two shoulders (at $T \approx 0.45$ and 1.0), all of which are statistically significant. As shown by a more detailed analysis [16], the shoulder at $T = 1$ is due to the collapse from open coil to molten globule, while that at $T = 0.45$ is due to the folding into the native state. There seems to be no state with $E = -97$, and very few states with $E = -96$ and -95, leading to an effective gap between $E = -94$ and $E = -98$. The main peak seems related to the formation of—mostly misfolded (helix-dominated)—secondary and tertiary structure. The low-temperature phase, however, contains mostly β sheets (see Fig. 2).

In summary, we showed that the pruned-enriched Rosenbluth method can be very effectively applied to protein structure prediction in simple lattice models. It is suited for calculating statistical properties and is very successful in finding native states. In all cases it did better than any previous MC method, and in many cases it found lower states than those which had previously been conjectured to be native. Especially, we have presented a new candidate for the native conformation of a four helix bundle sequence which had been studied before by several authors. We verified that ground states of the HP model are highly degenerate and have no gap, leading to bad folders. In contrast, the ground state of the four helix bundle sequence has a small gap and has low degeneracy because of the modified interaction strengths. But it folds only at very low T, and should not be a good folder either.

The authors thank G. Barkema, E. Domany, and M. Vendruscolo for discussions, and D. K. Klimov and R. Ramakrishnan for correspondence.

[1] *Protein Folding,* edited by T. E. Creighton (Freeman, New York, 1992).
[2] D. K. Klimov and D. Thirumalai, Proteins: Struct., Funct. Genet. **26**, 411 (1996); sequences are available from http://www.glue.umd.edu/klimov.
[3] K. A. Dill, Biochemistry **24**, 1501 (1985).
[4] K. F. Lau and K. A. Dill, Macromolecules **22**, 3986 (1989); J. Chem. Phys. **95**, 3775 (1991); D. Shortle, H. S. Chan, and K. A. Dill, Protein Sci. **1**, 201 (1992).
[5] N. D. Socci and J. N. Onuchic, J. Chem. Phys. **101**, 1519 (1994).
[6] E. O'Toole and A. Panagiotopoulos, J. Chem. Phys. **97**, 8644 (1992).
[7] P. Grassberger, Phys. Rev. E **56**, 3682 (1997).
[8] M. N. Rosenbluth and A. W. Rosenbluth, J. Chem. Phys. **23**, 256 (1955).
[9] J. Batoulis and K. Kremer, J. Phys. A **21**, 127 (1988).
[10] T. Garel and H. Orland, J. Phys. A **23**, L621 (1990).
[11] B. Velikson, T. Garel, J.-C. Niel, H. Orland, and J. C. Smith, J. Comput. Chem. **13**, 1216 (1992).
[12] C. J. Umrigar, M. P. Nightingale, and K. J. Runge, J. Chem. Phys. **99**, 2865 (1993).
[13] F. T. Wall and J. J. Erpenbeck, J. Chem. Phys. **30**, 634 (1959); *ibid.*, 637 (1959).
[14] H. Frauenkron and P. Grassberger, J. Chem. Phys. **107**, 9599 (1997).
[15] U. Bastolla and P. Grassberger, J. Stat. Phys. **89**, 1061 (1997).
[16] U. Bastolla, H. Frauenkron, E. Gerstner, P. Grassberger, and W. Nadler, Proteins (to be published).
[17] K. Yue *et al.*, Proc. Natl. Acad. Sci. USA **92**, 325 (1995).
[18] K. A. Dill, K. M. Fiebig, and H. S. Chan, Proc. Natl. Acad. Sci. USA **90**, 1942 (1993).
[19] K. Yue and K. A. Dill, Phys. Rev. E **48**, 2267 (1993).
[20] R. Unger and J. Moult, J. Mol. Biol. **231**, 75 (1993).
[21] R. Ramakrishnan, B. Ramachandran, and J. F. Pekney, J. Chem. Phys. **106**, 2418 (1997).
[22] N. D. Socci and J. N. Onuchic, J. Chem. Phys. **103**, 4732 (1995).
[23] A conformation with $E = -95$ containing also mainly β sheets was found in J. M. Deutsch, J. Chem. Phys. **106**, 8849 (1997), by means of an algorithm similar to that in [21], after one week of CPU time on a Pentium processor.
[24] A. M. Ferrenberg and R. H. Swendsen, Phys. Rev. Lett. **61**, 2635 (1988); **63**, 1195 (1989).